

# Symmetry and designability for lattice protein models

Tairan Wang<sup>1,2</sup>, Jonathan Miller<sup>1</sup>, Ned S. Wingreen<sup>1</sup>, Chao Tang<sup>1\*</sup>, and Ken A. Dill<sup>3</sup>

<sup>1</sup> *NEC Research Institute, 4 Independence Way, Princeton, New Jersey 08540*

<sup>2</sup> *Department of Physics, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139*

<sup>3</sup> *Department of Pharmaceutical Chemistry, University of California, San Francisco, California 94143*

Native protein folds often have a high degree of symmetry. We study the relationship between the symmetries of native proteins, and their *designabilities* – how many different sequences encode a given native structure. Using a two-dimensional lattice protein model based on hydrophobicity, we find that those native structures that are encoded by the largest number of different sequences have high symmetry. However only certain symmetries are enhanced, *e.g.*  $x/y$ -mirror symmetry and  $180^\circ$  rotation, while others are suppressed. If it takes a large number of mutations to destabilize the native state of a protein, then, by definition, the state is highly designable. Hence, our findings imply that insensitivity to mutation implies high symmetry. It appears that the relationship between designability and symmetry results because protein substructures are also designable. Native protein folds may therefore be symmetric because they are composed of repeated designable substructures.

## I. INTRODUCTION

The folded structures of proteins are often highly ordered. They are comprised of secondary structures, and often have striking regularities in their tertiary organization<sup>1,2</sup>. What is the origin of symmetry in natural proteins?

We approach this question by exploring the symmetries in simple lattice models of protein folding. Lattice models for proteins have been a rich source of information on protein structure. Yue and Dill<sup>3</sup> observed certain protein-like secondary structures and tertiary symmetries in HP lattice model proteins that have low degeneracies, *i.e.*, a small number of low energy states. More recently, Li et al.<sup>4</sup> noticed that the most “designable” structures, namely those with a large number of sequences folding into them, also often have global symmetries. Since the most designable structures also have other protein-like properties – they have sharp thermal folding transitions and are fast folders<sup>5</sup>, the connection to symmetry is intriguing.

In these earlier studies, no quantitative measure was used to define symmetry. Here, we explore in detail the connection between designability and global symmetry, based on a quantitative, but simple, measure of symmetry. Within the hydrophobic model<sup>6</sup>, we quantify the relation between designability and symmetry for  $6\times 6$  compact lattice proteins.

This article is organized as follows: Sec. II reviews the hydrophobic model and the designabilities of structures. In Sec. III, we relate symmetry to designability and identify the importance of the surface-core pattern to the particular emerging symmetries. To understand the origin of enhanced symmetry, in Sec. IV we explore, first, the role of surface-to-core transitions and, second, the extent to which symmetric folds result from the repeated use of designable substructures. For comparison, Sec. V addresses symmetry in a model not based on hydrophobicity. Sec. VI is the summary and conclusion.

## II. HYDROPHOBIC MODEL

In this section, we review the hydrophobic model and the designabilities of structures. For more details, the reader is referred to Li et al.<sup>6</sup>.

The hydrophobic model is a combination of the HP model<sup>7</sup> and the solvation model citeEisenberg86. In an HP model, the twenty different amino acids of proteins are replaced by two monomer types, Hydrophobic or Polar, according to their affinities for water. Each protein is therefore a sequence of H's and P's. In a *lattice* HP model, the amino acids are restricted to fall only on the sites of a regular lattice, typically a square lattice in two dimensions or a cubic lattice in three dimensions. The allowed conformations are self-avoiding, and hence cannot visit a single lattice site more than once.

Here we use a variant that we call the hydrophobic model, in which only the maximally compact structures are considered as possible ground states. This simplification still allows us to capture the essence of the HP model, but

---

\*Corresponding author; electronic mail: tang@research.nj.nec.com

with two advantages: a substantial reduction in computational cost, and a conceptually useful method to represent sequences and structures within the same kind of abstract spatial representation, described below. In the hydrophobic model, the energy of a compactly folded protein is taken to be simply minus the number of H monomers in the “core” (cf. Fig. 1). Therefore, in the hydrophobic model, the energy of an HP-protein sequence folded into a particular compact structure depends only on the structure’s ordering of surface and core sites. Thus, a structure can be represented by a string  $\mathbf{s}$  of 0’s and 1’s: sites in the core region are represented by 1’s and sites on the surface are represented by 0’s, as illustrated in Fig. 1. Sequences are also represented by strings of 0’s (P) and 1’s (H),  $\mathbf{h} = (h_1, h_2, \dots, h_N)$ , where  $h_i$  denotes the hydrophobicity of the monomer at position  $i$  of the sequence. The energy of a sequence folded into a particular structure is therefore given by

$$H = - \sum_{i=1}^N s_i h_i$$

where  $s_i$  is the structure string. An equivalent way of writing the energy is,

$$H = \frac{1}{2} \sum_{i=1}^N (s_i - h_i)^2 - \frac{1}{2} \sum_{i=1}^N s_i^2 - \frac{1}{2} \sum_{i=1}^N h_i^2.$$

The number of core sites is the same for all structures of the same size, thus  $\frac{1}{2} \sum s_i^2$  is a constant and can be dropped. Similarly, the last term,  $\frac{1}{2} \sum h_i^2$ , is a constant for each sequence and so does not influence which structure is the ground state for that sequence. Therefore, the only relevant term is the first term, which measures the Hamming distance between the structure string and the sequence string in an  $N$ -dimensional Euclidean space. A sequence with string  $\mathbf{h}$  will have a particular structure with string  $\mathbf{s}$  as its unique ground state if and only if  $\mathbf{h}$  is closer to  $\mathbf{s}$  than to any  $\mathbf{s}'$  corresponding to another structure.

The designability of a structure can therefore be obtained from the following geometric construction: Draw bisector planes between  $\mathbf{s}$  and all of its neighboring structures in the  $N$ -dimensional space. The volume enclosed by these planes is called the Voronoi polytope around  $\mathbf{s}$ . The designability of a structure is the number of sequences lying entirely within the Voronoi polytope around that structure. This is schematically represented in Fig. 2. Each vertex represents a sequence. Those vertices corresponding to structures are circled. Intuitively, the designability of a structure is closely related to how far away its nearest neighbors are. The further away its neighbors are, the more designable it is.

The histogram of the number of structures versus designability for the 6x6 hydrophobic model is shown in Fig. 3. The distribution has a long tail of highly designable structures compared to a Poisson distribution with the same mean. If sequences were randomly assigned to structures, the resulting distribution of designabilities would be Poisson. It is clear from Fig. 3 that the structures in the tail have anomalously high designabilities. That is, they are unique ground states of many more than their share of sequences.

The many sequences that have a particular highly designable native structure are related to each other by point mutations<sup>6</sup>. For the model we consider, a point mutation is simply the replacement of a hydrophobic monomer “1” by a polar monomer “0”, or vice versa. Often, many monomers can be independently mutated without destabilizing the native state<sup>4</sup>. Therefore, the folding of these sequences is relatively insensitive to mutations. One can think of highly designable structures as those which remain most stable under sequence mutations.

### III. SYMMETRY AND DESIGNABILITY

In Li et al.<sup>4</sup>, it was noted that highly designable structures tend to be highly symmetric, with global mirror symmetries as well as regular local motifs. In this study, we explore the connection between designability and symmetry in detail. We focus on 6x6 2D square-lattice proteins.

To measure the symmetry of a structure, we look at how well that structure is preserved under rigid global transformations. Specifically, the transformations are the mirror reflections about the x/y axes, the mirror reflections about the two diagonal directions, and 90° and 180° rotations. The symmetry scores for a given structure are the number of overlapping bonds between that structure and each of its transformed versions. The maximum possible symmetry score for a 6x6 compact structure is 35.

### A. Hydrophobic Model with Centered Core

We begin by studying the trends of symmetry versus designability for the hydrophobic model. The symmetry scores, averaged over designability bins, are plotted versus the designability in Fig. 4. It is observed that, on average, the x/y-mirror symmetry (the larger of the x-mirror symmetry score and the y-mirror symmetry score) increases with designability (Fig. 4(a)). A similar trend is observed for  $180^\circ$  rotation symmetry, which is consistent with the x/y-mirror symmetry result since a  $180^\circ$  rotation is simply an x-mirror operation followed by a y-mirror operation. For the other symmetry operations,  $90^\circ$  rotation and diagonal mirrors, the trend is reversed – higher designability implies lower symmetry scores for these symmetries. Thus, for the hydrophobic model, there is indeed a connection between designability and symmetry as previously noted<sup>4</sup>. However, different symmetries behave differently with increasing designability. In this case, the x/y-mirror symmetry and  $180^\circ$  rotation symmetry are enhanced for highly designable structures.

### B. Hydrophobic model with shifted core

For the hydrophobic model, the surface-core pattern has the symmetry of a square (Fig. 1). What if this is not the case? Does higher designability always lead to higher x/y-mirror symmetry scores, even when the surface-core pattern is disrupted? To address this question, we study a shifted-core version of the hydrophobic model. The core sites have been shifted to the lower corner (Fig. 5). In this shifted-core model, the energy of a compactly folded protein is taken to be simply minus the number of H monomers in the new off-center “core”<sup>9</sup>. The histogram of designability for the shifted-core model is shown in Fig. 6. Again, the region of high designability is characterized by a long tail, qualitatively similar to that of the ordinary 6x6 model.

With the “core” shifted to one corner, the surface-core pattern no longer has mirror symmetries about the x and y axes. In fact, only one diagonal mirror symmetry is left. In Fig. 7, we plot the averaged symmetry scores versus designability for the shifted-core model. When x/y-mirror symmetry is plotted versus designability there is no correlation. Instead, only the diagonal-mirror symmetry which is present in the surface-core pattern increases significantly with designability (Fig. 7(a)).

These results indicate that the surface-core pattern is important in determining which symmetries are favored in highly designable structures. In both cases considered, the preferred symmetries follow the surface-core pattern.

## IV. ORIGIN OF SYMMETRY

Why is there an enhancement of symmetry for highly designable structures? Also, why are some symmetries enhanced and not others? In this section, we examine two possible origins of symmetry. First, perhaps global symmetries arise in designable structures because of a high number of surface-to-core transitions. Or second, perhaps designable structures have global symmetries because they arise from repeated highly designable substructures.

### A. Surface-to-core transitions

A possible candidate for the link between designability and symmetry is a local property of structures – the number of surface-to-core transitions. A surface-to-core transition occurs when monomer  $i$  of the chain is in the core and monomer  $i + 1$  is on the surface, or vice versa. Previously<sup>6,10</sup>, it was observed that highly designable structures have an excess of surface-to-core transitions. The connection can be understood as follows: (1) A structure with a large number of surface-to-core transitions is difficult to rearrange without exchanging many surface and core sites. Such structures are therefore likely to be far from their neighbors in the space of strings, and thus have a chance for high designability (cf. Fig. 2). (2) In turn, a structure with a large number of surface-to-core transitions has a geometrical regularity which may naturally lead to global symmetry. Moreover, the geometrical regularities, and hence the enhanced global symmetry, should reflect the symmetry of the surface-core pattern, consistent with our results using the shifted-core model.

We tested whether surface-to-core transitions form the link between designability and global symmetry. We find that, qualitatively, both correlations (1) and (2) are present, however, quantitatively, they fail to account for the observed enhancement of global symmetry. To quantify the first correlation, the number of surface-to-core transitions averaged over structures of a given range of designabilities is plotted against designability in Fig. 8(a) for the original

hydrophobic model. High designability clearly implies an enhanced number of surface-to-core transitions. To demonstrate the second correlation, the x/y symmetry score averaged over structures with a given number of surface-to-core transitions is plotted against the number of surface-to-core transitions in Fig. 8(b). Symmetry does increase with the number of surface-to-core transitions when the number of transitions is large<sup>11</sup>.

Is the chain of correlations from designability to surface-to-core transitions to global symmetry strong enough to explain the observed enhancement of global symmetry? In Fig. 8(c), the x/y-symmetry score is plotted against designability assuming the connection between them is only through the correlation of each with the number of surface-to-core transitions. Specifically, for a given designability, the corresponding average number of surface-to-core transitions is obtained from panel (a), then the corresponding average x/y-symmetry score for that number of surface-to-core transitions is obtained from panel (b). The predicted x/y-symmetry score thus obtained is then plotted against designability. The actual x/y-symmetry score versus designability from Fig. 4(a) is also plotted. We see that surface-to-core transitions account for only a fraction of the observed connection between symmetry and designability.

## B. Designable substructures

A second possible explanation for why designable folds are so symmetric is (1) they arise from designable substructures, and (2) symmetries are a natural consequence of assembling anything from identical substructures.

The most designable structure for the 6x6 hydrophobic model is shown in the left half of Fig. 9. We take the right half of the 6x6 surface-core pattern (a 3x6 rectangle) and calculate the designabilities of all possible structures for this 3x6 hydrophobic model. The most designable 3x6 structure is shown in the right half of Fig. 9. Comparing the two structures, we see that the most designable 3x6 structure is very similar to one half of the most designable 6x6 structure of the original model. We conclude that this 6x6 structure is highly designable because it is composed of two highly designable substructures. The role of symmetry in this case can then be understood as duplicating a winning solution.

It is not yet clear how to quantify this concept of designable substructures. Any scheme that involves breaking and reforming bonds, as would be necessary to relate the structures in Fig. 9, seems arbitrary and unsatisfactory. Nevertheless, a connection between designable components and global symmetry seems to us likely, and may have implications for understanding global symmetries in real proteins.

## V. SYMMETRY BEYOND THE HYDROPHOBIC MODEL

As a final question, we ask if the connection between designability and symmetry is particular to models based on hydrophobicity, or whether it occurs more generally. In the hydrophobic model each structure is characterized by its ordering of surface and core sites. As an alternative, we consider a model in which each structure is characterized by its complete contact matrix, as described below. Those structures with large contact-matrix distance from their neighbors are considered to be highly designable. Within the contact-matrix model, the designable structures *do not* show significantly enhanced symmetry.

Each structure has a contact matrix for its monomers. The elements of the contact matrix between monomers are 1 if they are next to each other in the structure, but not adjacent on the chain, and 0 otherwise. A compact structure is uniquely defined by its contact matrix, up to rigid rotations and inversion. Thus, contact matrices and structures are related by a one-to-one mapping.

The distance between any two structures can be measured by the overlap between their contact matrices. The more overlap, the more bonds they have in common. Hence, contact-matrix distance measures the similarity of structures without particular emphasis on surface and core sites. Just as in the hydrophobic model, where structures with few neighboring strings emerged as highly designable, structures with few neighbors in contact-matrix distance would emerge as highly designable for more general models of amino-acid interaction<sup>12</sup>. For the set of compact 6x6 structures, we take the number of neighbors within a contact-matrix distance of 16 as a measure designability, with *low* number of neighbors implying *high* designability. The histogram of number of structures versus number of near neighbors is shown in Fig. 10(a). Shown in Fig. 10(b) is a plot of the averaged top symmetry scores versus the number of neighboring structures within the cutoff distance of 16. The top symmetry score for a structure is the highest score for all possible rotations and reflections. The horizontal line indicates the average top symmetry score of 27.7. The region of few neighbors, and hence high designability, is at the left of the figure and has only a very slightly enhanced symmetry score with respect to the average.

We conclude that enhanced global symmetry of designable structures *does not* emerge generally from models with arbitrary interactions among amino acids. Rather, the enhancement of global symmetries is particular to models in

which the interaction between amino acids is dominated by hydrophobicity<sup>4,6</sup>. It appears that the correlation between the designability and symmetry of a native protein is a consequence of the key role played by hydrophobic solvation, and the approximate radial symmetries that result from it.

## VI. CONCLUSION

In this work, we have examined the connection between the designability and symmetry of protein structures within the hydrophobic model of Li et al.<sup>6</sup>. The designable structures, namely those which are unique ground states of many more than their share of sequences, had been previously identified to have enhanced global symmetry, as well as other protein-like attributes such as thermodynamic stability and stability against mutations<sup>4</sup>. To quantify the relation between symmetry and designability we focused on the set of two-dimensional compact structures which fill the sites of a 6x6 square lattice. We found that the designable structures have strongly enhanced symmetry for x/y reflection and 180° rotation. For a related model in which the “core” is shifted to one corner, the only enhanced symmetry was a diagonal reflection. This indicates that the enhanced symmetry of the designable structures reflects a symmetry of the surface-core pattern.

To explore the origin of symmetry, we examined the relation between designability, number of surface-to-core transitions, and global symmetry. We conclude that an increase in surface-to-core transitions among designable structures can only account for a fraction of the observed enhancement of global symmetry. Our working hypothesis is that the global symmetry of designable structures results from the repetition of designable substructures.

Finally, from a comparison model based on contact-matrix distances we conclude that the relation between designability and symmetry originates from surface-core symmetries, which in turn, result mainly from hydrophobic interactions.

- 
- <sup>1</sup> M. Levitt and C. Chothia, *Nature* **261**, 552 (1976).
- <sup>2</sup> J. S. Richardson, *Proc. Natl. Acad. Sci. USA* **73**, 2619 (1976); *Adv. Protein Chem.* **34**, 167 (1981).
- <sup>3</sup> K. Yue and K. A. Dill, *Proc. Natl. Acad. Sci. USA* **92**, 146 (1995).
- <sup>4</sup> H. Li, R. Helling, C. Tang, and N. Wingreen, *Science* **273**, 666 (1996).
- <sup>5</sup> R. Mélin, H. Li, N. S. Wingreen, and C. Tang, *J. Chem. Phys.* **110**, 1252 (1999).
- <sup>6</sup> H. Li, C. Tang, and N. S. Wingreen, *Proc. Natl. Acad. Sci. USA* **95**, 4987 (1998).
- <sup>7</sup> K. A. Dill, *Biochemistry* **24**, 1501 (1985); K. F. Lau and K. A. Dill, *Macromolecules* **22**, 3986 (1989).
- <sup>8</sup> D. Eisenberg and A. D. McLachlan, *Nature* **319**, 199 (1986).
- <sup>9</sup> The shifted-core model is not intended to be physically realistic. It is, however, useful to probe the connection between symmetry and designability.
- <sup>10</sup> C. T. Shih, Z. Y. Su, J. F. Gwan, B. L. Hao, C. H. Hsieh, and H. C. Lee, *Phys. Rev. Lett.* **84**, 386 (2000).
- <sup>11</sup> Symmetry is also enhanced for structures with minimal numbers of surface-to-core transitions, but these represent only a small subset of the many structures with low designability.
- <sup>12</sup> N. E. G. Buchler and R. A. Goldstein, *J. Chem. Phys.* **112**, 2533 (2000).

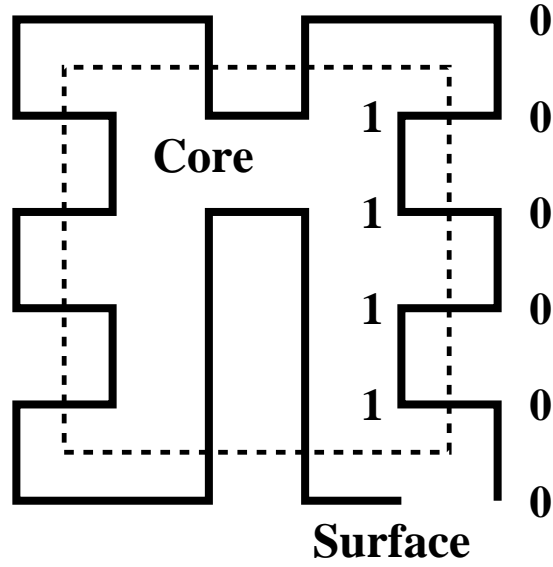


FIG. 1. The most designable structure in the 6x6 hydrophobic model. The 16 sites in the core region, enclosed by the dashed lines, are represented by 1's; the 20 sites on the surface are represented by 0's. Hence the structure is represented by the string 001100110000110000110011000011111100. The structure has an approximate mirror symmetry, with an x-mirror symmetry score of 34, i.e. with 34 bonds superposing upon reflection. The structure is also highly “pleated” with 12 surface-to-core transitions.

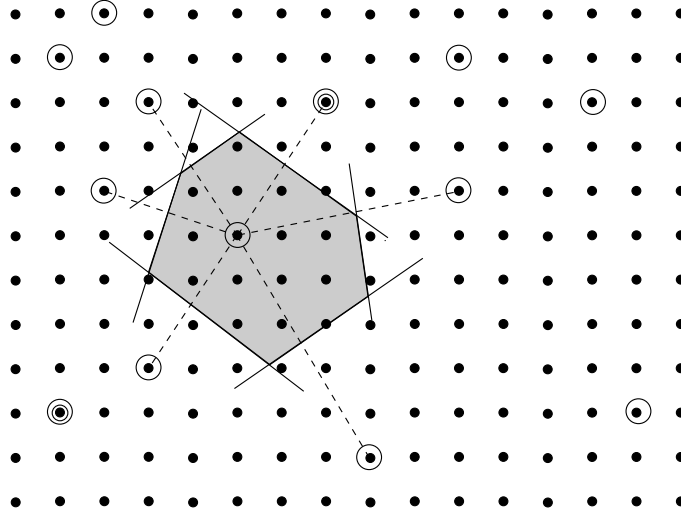


FIG. 2. Schematic representation of sequences and structures as binary strings. Each vertex represents a possible sequence, *i.e.* a string of 36 0's and 1's. Those vertices corresponding to structures are circled. The sequences lying closer to a particular structure than to any other have that structure as their unique ground state. The designability of a structure is therefore the number of sequences lying entirely within the Voronoi polygon about that structure. In cases where more than one structure has the same string of 0's and 1's, *i.e.*, the same pattern of surface-core sites, the corresponding vertices are circled twice. These structures have zero designability.

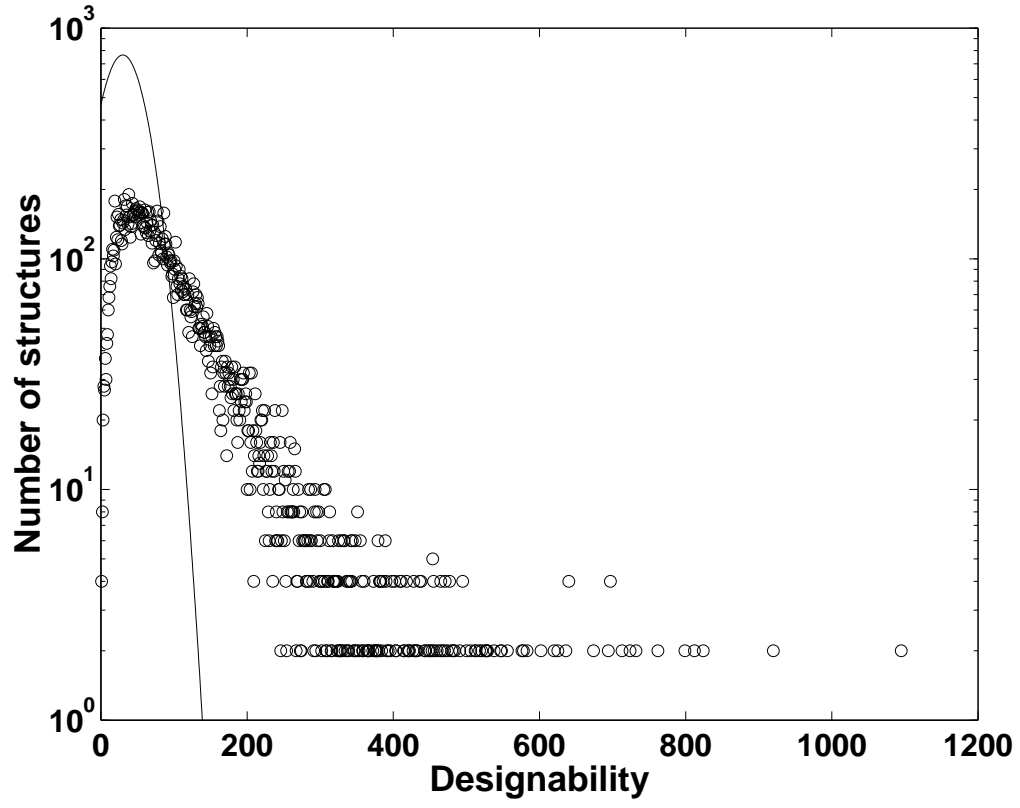


FIG. 3. Histogram of the number of structures versus designability for the 6x6 hydrophobic model. The data is generated by sampling  $20 \times 10^6$  sequence strings. For comparison, the solid line shows the Poisson distribution with the same average designability.

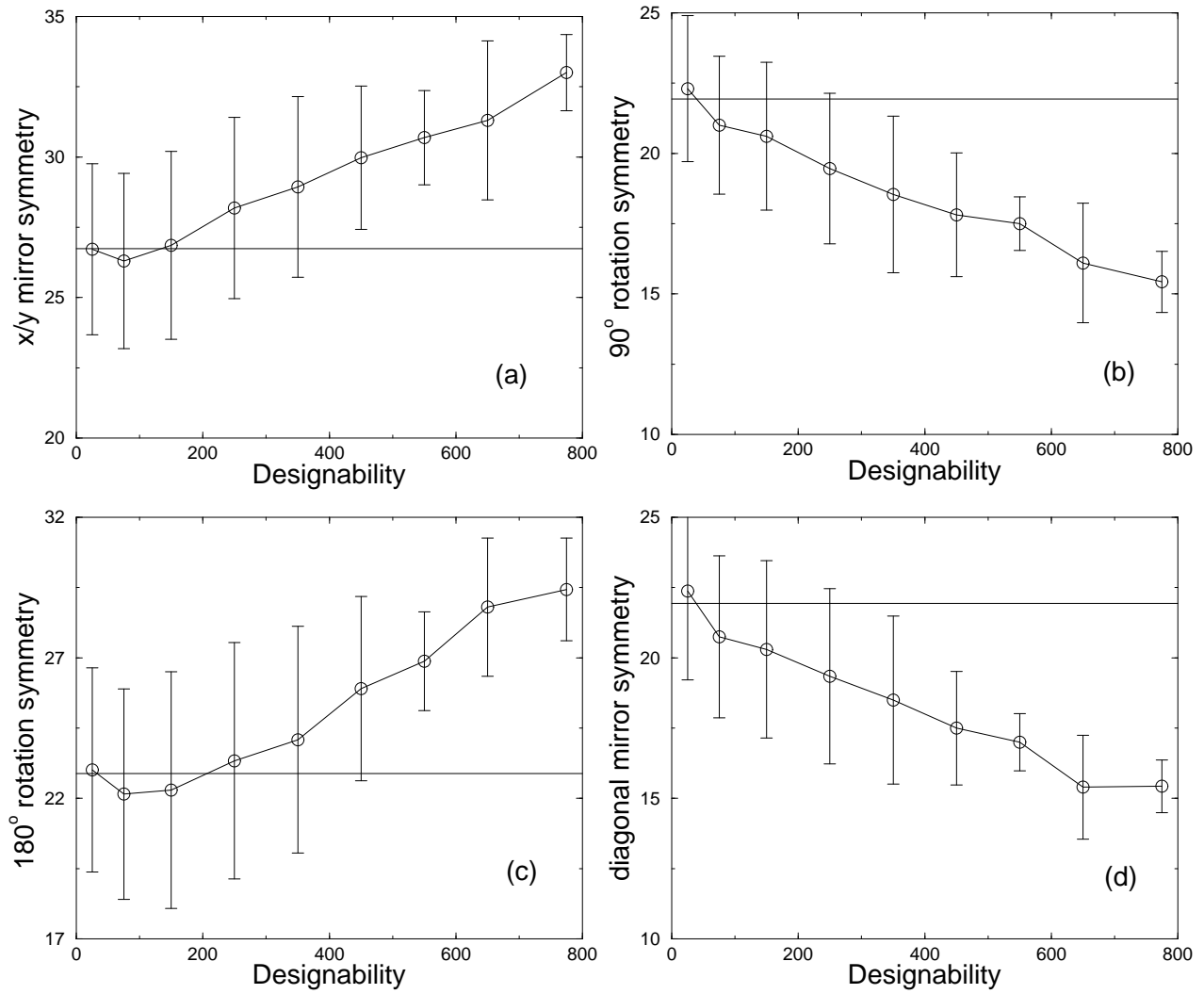


FIG. 4. Averaged symmetry scores versus designability for the hydrophobic model. The data is collected into bins according to designability. The circles are the average symmetry scores within a designability bin, and the error bars are the standard deviations. The horizontal lines indicate the overall averaged symmetry scores in each case. The x/y-mirror symmetry and the  $180^\circ$  rotation symmetry increase with designability, as shown in panels (a) and (c). The other symmetries decrease with designability, as shown in panels (b) and (d).

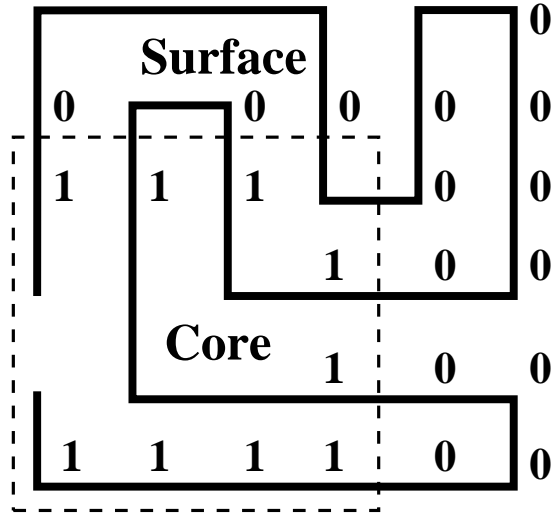


FIG. 5. Shifted-core hydrophobic model. The core region is shifted to the lower left. Only one diagonal symmetry is still present in the surface-core pattern.

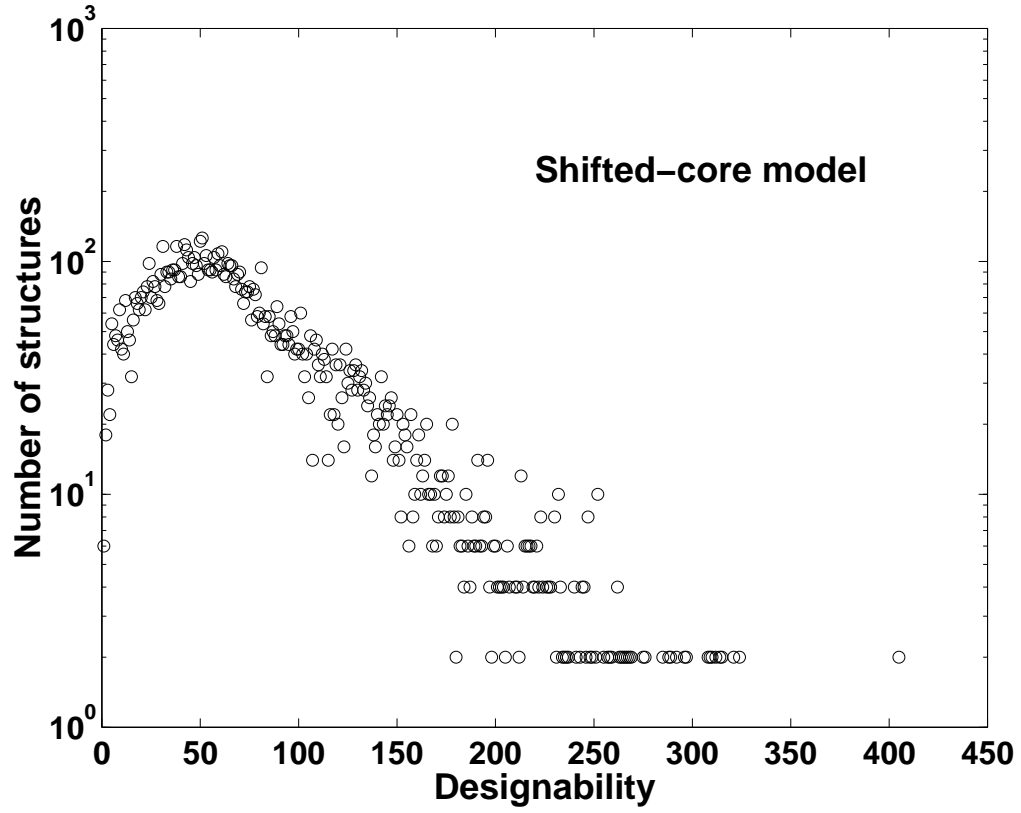


FIG. 6. Histogram of the number of structures versus designability for the shifted-core model. The data is generated from a random sampling of  $6.3 \times 10^6$  sequence strings.

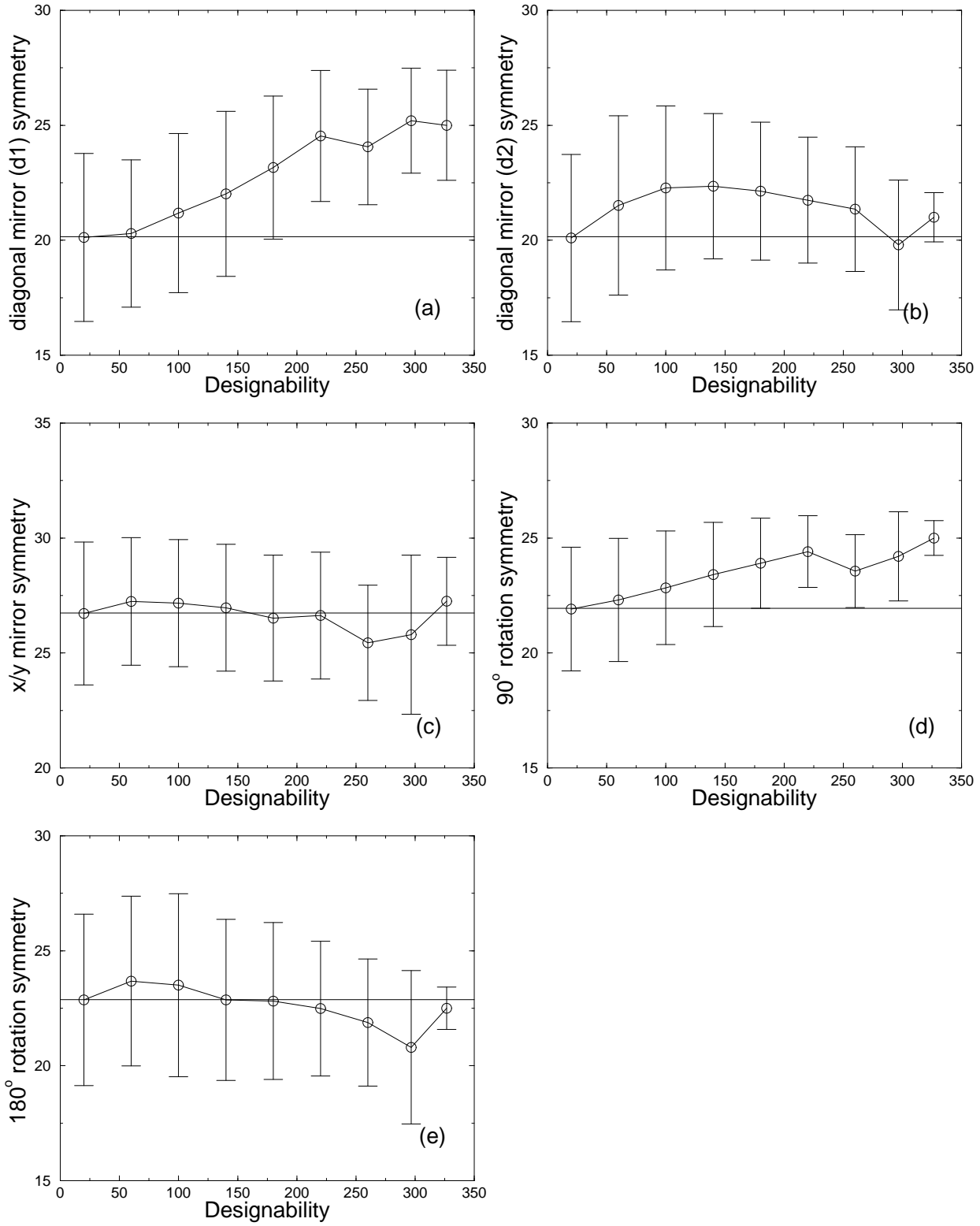


FIG. 7. Averaged symmetry numbers versus designability for the shifted-core model. “d1” is the diagonal-mirror symmetry preserved in the surface-core pattern; “d2” is the other diagonal-mirror symmetry. The horizontal lines indicate the overall average symmetry scores in each case. The diagonal-mirror symmetry “d1” increases with designability, while the other symmetries show little change.

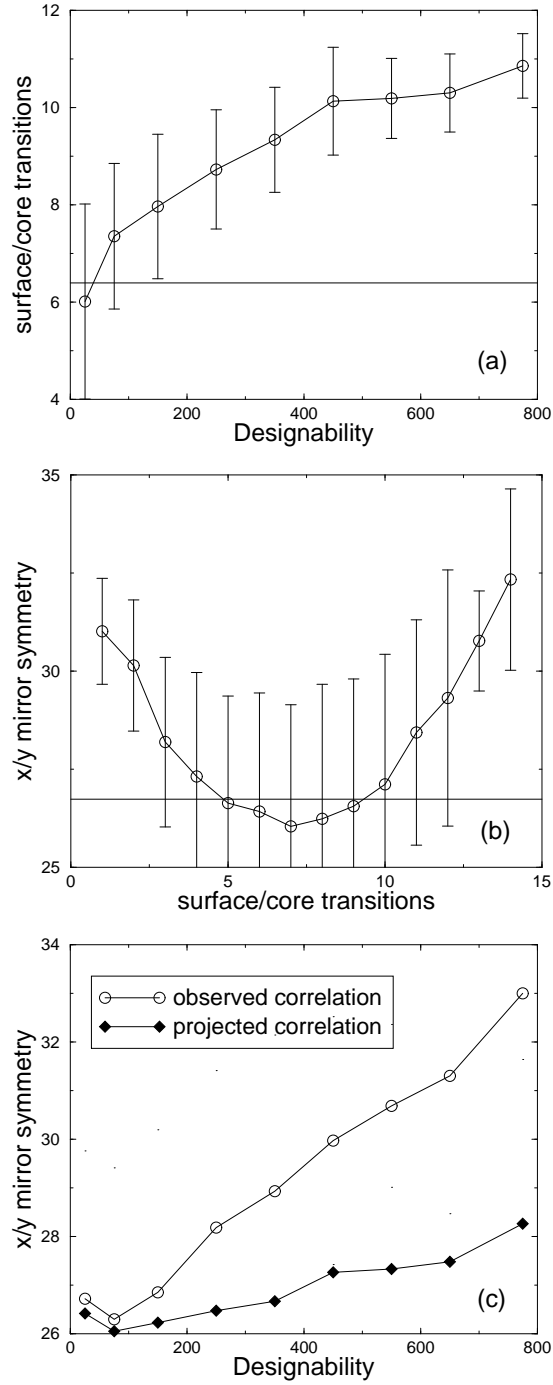


FIG. 8. Panel (a) – Averaged number of surface-to-core transitions versus designability. The horizontal line indicates the overall averaged number of surface-to-core transitions. Panel (b) – Averaged x/y-symmetry score versus number of surface-to-core transitions. The horizontal line indicates the overall averaged x/y-symmetry score. Panel (c) – Open circles indicate the averaged x/y-symmetry score versus designability; filled diamonds give the predicted x/y-symmetry score versus designability if we assume the connection between symmetry and designability is only through the correlation of each with the number of surface-to-core transitions.

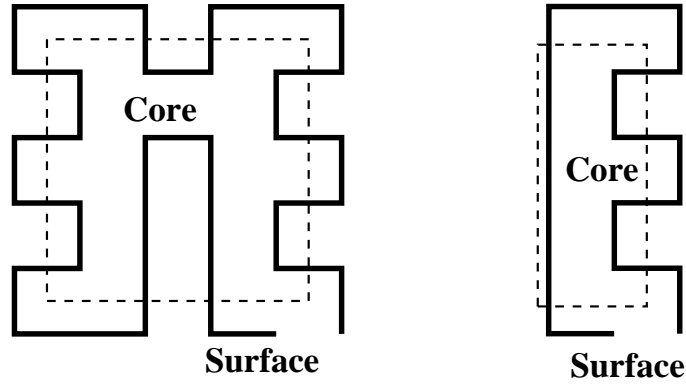


FIG. 9. The left half shows the most designable structure for the 6x6 hydrophobic model; the right half shows the most designable structure for a 3x6 hydrophobic model which corresponds to one half of the 6x6 model.

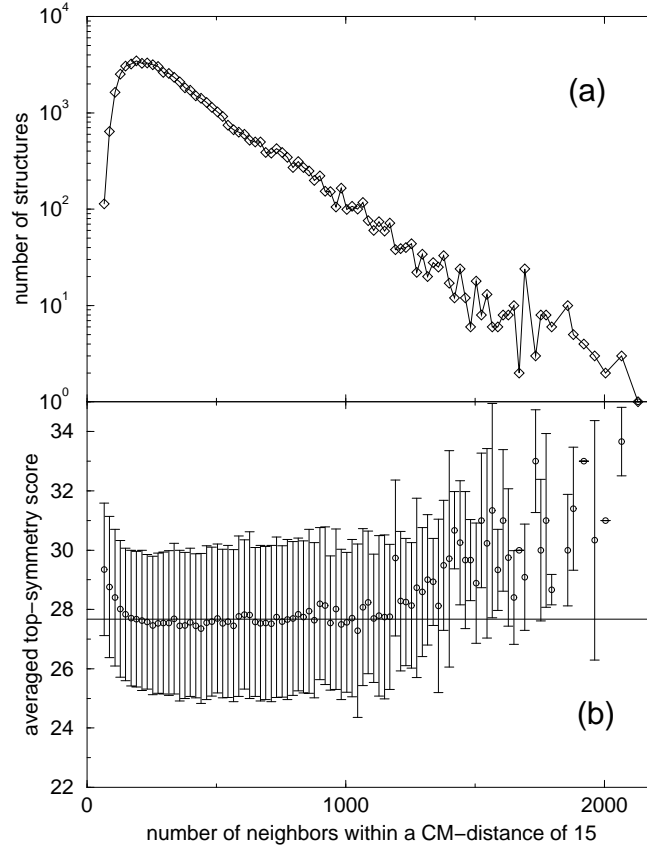


FIG. 10. Panel (a) – Histogram of the number of structures versus number of neighboring structures within a Contact-Matrix (CM) distance of 15. The maximum CM distance between any two structures is 25. Panel (b) – Averaged top-symmetry scores.